Neighboring Base Effects on Substitution Rates in Pseudogenes[1]

Michael Bulmer
Department of Biomathematics, University of Oxford

Substitution rates in pseudogenes can be used to estimate the frequencies of different types of mutation on the assumption that pseudogenes are not subject to selective constraints. These rates are used here to investigate the effect of neighboring bases on mutation rates. There is a marked increase in the frequency of transitions, though not of transversions, from the doublet CG. There are also some smaller effects of neighboring bases on the frequencies of transitions from adenine and thymine. The results are used to predict dinucleotide frequencies in a stretch of DNA subject to no selective constraints and to investigate the possibility of non-randomness in the usage of stop codons.

Introduction

Since pseudogenes are apparently functionless (Li et al. 1981), substitution rates in them should be proportional to mutation rates—and can be used to estimate the relative frequencies of different types of mutation. Gojobori et al. (1982) and Li et al. (1984) have used this fact to estimate the pattern of point mutations in nucleotides and have observed a high frequency of transitions from CG dinucleotides, which can be explained by the observation that at this doublet in vertebrates there is a high level of methylated cytosine that mutates abnormally frequently to thymine (Coulondre et al. 1978; Bird 1980; Razin and Riggs 1980). The purpose here is to investigate systematically the effect of neighboring bases on substitution rates in pseudogenes in order to estimate the magnitude of the CG effect and to determine whether other neighbor effects exist.

Material and Methods

The following 14 vertebrate pseudogenes were studied: human beta-tubulin ψ7 and ψ14 (Lee et al. 1983); *Xenopus* 5S rRNA (Jacq et al. 1977; Miller et al. 1978); human U1 small nuclear RNA ψ1, 7, 8, 11, 101 (Denison et al. 1981; van Arsdell et al. 1981; Monstein et al. 1983; Lund and Dahlberg 1984); human immunoglobulin V-kappa (Bentley and Rabbitts 1980; Nishioka and Leder 1980); human immunoglobulin C-epsilon ψ3 (Flanagan and Rabbitts 1982; Max et al. 1982; Ueda et al. 1982; Hisajima et al. 1983); human and mouse alpha-globin (Proudfoot et al. 1977; Heindell et al. 1978; Nishioka and Leder 1979; Michelson and Orkin 1980; Nishioka et al. 1980; Proudfoot and Maniatis 1980); rabbit and goat ψ^X beta-globin (Hardison et al. 1979; Konkel et al. 1979; Lacy and Maniatis 1980; Lawn et al. 1980; Cleary et al. 1981; Schon et al. 1981).

1. Key words: pseudogene, substitution rate, methylated cytosine, dinucleotide frequency, stop codon.

Address for correspondence and reprints: Dr. Michael Bulmer, Department of Biomathematics, University of Oxford, 5 South Parks Road, Oxford OX1 3UB, England.

Mol. Biol. Evol. 3(4):322–329. 1986.

0737-4038/86/0304-3401$02.00

Sequences were taken from the GenBank and EMBL nucleic acid data banks. Each pseudogene was aligned with the relevant functional gene, using the published alignment allowing for insertions and deletions as required and considering only the coding region of the gene. To ensure that the direction of any observed change was known, only those sites were used at which one could be reasonably confident that no change had occurred in the functional gene since the origin of the pseudogene. The following criteria were used to this effect: (1) All sites were used for the two RNA genes, following Gojobori et al. (1982), since nucleotide substitution is so slow in such genes that it is safe to assume that no substitutions have occurred in the functional genes. (2) For beta-tubulin and immunoglobulin C-epsilon only those sites in the functional gene were used at which a nucleotide substitution must cause an amino acid replacement, following Li et al. (1984). The amino acid sequence is highly conserved in these genes, so that it is safe to assume that no change has occurred in the functional gene at these sites; but it would be unsafe to make this assumption with respect to the remaining sites at which synonymous substitutions are possible. (3) For the more variable immunoglobulin V-kappa and globin genes, only those sites in the functional gene were included at which the observed nucleotide was the same as the consensus nucleotide for the sequences of two or three species (man and mouse for the immunoglobulin V-kappa gene; man, mouse, and rabbit for the alpha- and beta-globin genes); if no consensus (i.e., strict majority) existed, the site was excluded.

No obvious evidence of gene conversion was seen in any of the pseudogenes studied. Even if it had been present it would not have affected the conclusions reached about the relative frequencies of different substitutions.

Results

Table 1 shows substitutions summed over all 14 pseudogenes and over both the coding and the noncoding strands. (A preliminary analysis showed no difference between the two strands, as expected a priori. For example, transitions from A to G on the coding strand are about as frequent as those from T to C, which correspond with the transitions from A to G on the noncoding strand.) Only the first half of the table (for A and C) is shown, since the full table is point symmetric about the center. The two types of transversion (C or T from A, A or G from C) are about equally frequent and will be pooled in future discussion. Note that transitions are more frequent than transversions and that substitutions are more frequent from C or G than from A or T, in line with previous results (Gojobori et al. 1982; Li et al. 1984).

To assess the effect of a neighboring nucleotide, it is necessary to ensure that that nucleotide has not changed; in addition to the exclusions at the site itself, mentioned above, any sites that were followed by insertions or deletions in the pseudogene or for which the neighboring nucleotide differed in the functional gene and the pseudogene

Table 1
Base Substitutions in 14 Pseudogenes

Base in Functional Gene	Base in Pseudogene: A	C	G	T	Transitions (%)	Transversions (%)
A	2,216	47	106	36	4.4	3.5
C	79	2,274	69	283	10.5	5.5

were also excluded. The effect of the base on the right on substitutions in the pseudogene is shown in table 2. As before, data have been summed over all 14 pseudogenes and over both the coding and the noncoding strands. It is not necessary to consider separately the effect of the base on the left since this is given by symmetry; for example, the number of transitions from A with G on the right is equal to the number of transitions from T with C on the left.

There is no evidence of any effect of the neighboring base on the frequency of transversions. There is a large neighbor effect on transitions from C, one that results entirely from an increased frequency with G on the right; note that CG is self-complementary, so that there is a similar increase in transitions from G with C on the left. There are also smaller neighbor effects, which are similar in kind, on transitions from A or T; the transition frequency from either of these bases is reduced by having G on the right (or C on the left) and is increased by having T on the right (or A on the left).

The CG effect probably results from the occurrence of methylated cytosine in this doublet in vertebrates and accounts for the deficiency of CG doublets (Bird 1980). Since there is variability in the degree of CG doublet deficiency between different parts

Table 2
The Effect of the Neighboring Base on Substitutions in 14 Pseudogenes

Base in Functional Gene	Base on Right	Transitions (%)	Transversions (%)	Total Bases
A	A	4.0	2.5	398
A	C	4.6	1.8	565
A	G	2.4	2.1	673
A	T	6.1	4.3	396
χ^2 for heterogeneity		9.5*	7.0 NS	
C	A	4.9	5.1	691
C	C	6.9	3.6	798
C	G	44.0	5.8	225
C	T	6.0	4.5	516
χ^2 for heterogeneity		331.7**	2.8 NS	
G	A	10.4	6.3	511
G	C	9.1	4.7	616
G	G	10.1	3.8	685
G	T	10.8	5.0	517
χ^2 for heterogeneity		1.0 NS	3.9 NS	
T	A	4.0	1.5	202
T	C	4.0	2.6	647
T	G	2.2	3.6	810
T	T	5.6	2.8	394
χ^2 for heterogeneity		9.3*	2.9 NS	

NOTE.—The χ^2 values for heterogeneity test for an effect of the base on the right. NS = not significant.
* $P < .05$ on 3 degrees of freedom.
** $P < .001$ on 3 degrees of freedom.

of the genome (Smith et al. 1985), it is of interest to see whether there is any variability in the magnitude of the CG effect between pseudogenes derived from different functional genes. Table 3 shows that there is no evidence of heterogeneity, but the data are somewhat inconclusive since immunoglobulin V-kappa and beta-globin are the only genes in the table with a large CG-doublet deficiency.

For simplicity, the analyses in tables 1 and 2 have been performed on pooled data. Since it is possible under some circumstances to obtain spurious results when pooling contingency tables, a more complicated analysis was done on the unpooled data using a logistic regression model with the statistical package GLIM (McCullagh and Nelder 1983). The conclusions were identical. In particular, there was no difference between the behavior of processed and nonprocessed pseudogenes.

The results in table 2 can be used to estimate substitution rates (relative to an arbitrary time scale) as follows: (1) For transversions, there is no evidence of a neighbor effect, and the substitution rate can be estimated as the overall proportion of substitutions—which is 0.028 from A or T and 0.046 from C or G—divided equally between the two types of transversion. (2) For transitions from C, the rate can be taken as 0.060 with A, C, or T on the right, since there is no evidence of heterogeneity between these three bases on the right. The frequency of transitions from C with G on the right is 0.440, but this figure should be increased to 0.580 (= $-\ln[1-0.44]$) in estimating the transition rate to allow for the nonlinearity in the graph of frequency of substitution plotted against time. (This calculation assumes that the proportion of unchanged CG doublets declines exponentially with time. The correction is unnecessary for the other, much smaller frequencies.) There is no effect of the base on the left on transitions from C, since there is no effect of the base on the right on transitions from G. By symmetry, the transition rate from G with A, G, or T on the left is 0.060, and with C on the left it is 0.580; there is no effect of the base on the right. (3) For transitions from A or T there is a smaller effect of the base on both sides. (Remember that an effect of the base on the right of A implies an effect of the complementary base on the left of T, and vice versa.) The results in table 2 can be approximately reproduced by taking the transition rate from A or T with A or C on the right and G or T on the left as 0.042 and multiplying this rate by 0.55 with G on the right or C on the left or by 1.38 with T on the right or A on the left.

Discussion

The main conclusion is that there is a tenfold increase in the frequency of transitions in CG doublets in vertebrates, an increase that can be attributed to the high

Table 3
CG Effect for Different Genes

Gene	No. of Pseudogenes	% Transitions from C (*N*)	
		G on Right	G Not on Right
Beta-tubulin	2	43.8 (73)	3.9 (620)
Xenopus 5S rRNA	1	33.3 (3)	2.4 (42)
U1 RNA	5	30.9 (58)	4.9 (325)
Immunoglobulin V-kappa	1	100.0 (2)	7.7 (104)
Immunoglobulin C-epsilon	1	46.4 (56)	7.6 (278)
Alpha-globin	2	60.6 (33)	7.2 (293)
Beta-globin	2	100.0 (2)	8.5 (343)

frequency of methylated cytosine in this doublet. There are also some smaller effects of neighboring bases on frequencies of transitions from A and T.

These neighbor effects will introduce nonrandomness into the sequence of nucleotides. Table 4 shows the predicted dinucleotide frequencies in a stretch of DNA with the substitution rates estimated above and subject to no selective constraints. It is AT rich, a characteristic observed in vertebrate noncoding regions (Gojobori et al. 1982). There is a reduction in CG doublets to about one fifth of the frequency expected under independence; there is an increase of ~25% above expectation in the frequencies of CA, TG, AG, and CT doublets. These predictions agree qualitatively with observed dinucleotide frequencies in vertebrates (Bird 1980; Nussinov 1984), though no information is available for noncoding regions. The main exception is that the observed deficiency of TA doublets is not predicted in table 4. A model incorporating the CG effect but ignoring the neighbor effect for A and T predicts a similar decrease in TG doublets and a similar increase in CA and TG doublets (which are reached from CG by a single transition) but a smaller increase (by 9%) in AG and CT doublets. (The dinucleotide frequencies were condensed from the equilibrium distribution for trinucleotides, which was obtained by using the theory of continuous-time Markov processes [Cox and Miller 1965]. To obtain numerical results by an iterative procedure it was assumed that dependencies in the sequence do not extend farther than two bases either way; this is not exactly true but should provide a satisfactory approximation.)

These results are also of value in interpreting codon usage. Consider as an example the usage of the three stop codons TAG, TAA, and TGA, which will be numbered 1, 2, and 3. If m_{ij} is the rate of mutation from stop codon i to stop codon j, the relative frequency, in the absence of selection, of stop codon i, p_i, satisfies the differential equation

$$\frac{dp_i}{dt} = \sum_{j \neq i} m_{ji} p_j - \sum_{j \neq i} m_{ij} p_i. \quad (1)$$

(Any mutation to a non–stop codon can be ignored, since it will be rapidly eliminated.) Setting these equations to zero and observing that $m_{13} = m_{31} = 0$, since TAG and TGA are not connected by a single step mutation, we find that the equilibrium frequencies should be

$$\begin{aligned} p_1 &= m_{21}\, m_{32}/k, \\ p_2 &= m_{12}\, m_{32}/k, \\ p_3 &= m_{12}\, m_{23}/k, \\ (k &= m_{21}\, m_{32} + m_{12}\, m_{32} + m_{12}\, m_{23}). \end{aligned} \quad (2)$$

From the results at the end of the last section, the mutation rates in vertebrates are estimated as

$$\begin{aligned} m_{12} &= m_{32} = 0.06, \\ m_{21} &= 0.056, \\ m_{23} &= 0.040. \end{aligned} \quad (3)$$

The equilibrium frequencies of the three stop codons in vertebrates should therefore be

Table 4
Predicted Dinucleotide Frequencies per Thousand

	Base on Right				
Base on Left	A	C	G	T	Total
A	86	66	74	86	312
C	74	34	7	74	188
G	54	34	34	66	188
T	98	54	74	86	312
Total	312	188	188	312	1,000

$$p_1 = 0.36,$$
$$p_2 = 0.38, \quad (4)$$
$$p_3 = 0.26.$$

Table 5 shows the observed stop codon usage in a number of species. Only the Epstein-Barr virus seems to have the predicted random usage. All the other species avoid TAG, some having a preference for TGA, others for TAA. These facts suggest that selective pressures act on stop codon usage as they do on the usage of amino acid codons (Ikemura 1985). Two possible selective factors are (1) differential recognition by the release factors that catalyze termination and (2) differential misreading by tRNAs leading to readthrough. However, at least part of the preference for TGA in vertebrates may be due to the general tendency to avoid the doublet TA, a tendency that is not accounted for by mutation rates derived from table 2.

Table 5
Stop Codon Usage (%) in Different Species

	Stop Codon Usage (%)		
Species (*n*)	TAG	TAA	TGA
Man (115)	17	29	55
Mouse (89)	19	26	55
Chicken (39)	15	44	41
Drosophila (27)	15	85	0
Yeast (40)	12	70	18
E. coli (154)	6	71	23
Phage lambda (60)	7	37	56
Phage T7 (56)	11	55	34
Epstein-Barr virus (58) ...	36	38	26

Acknowledgment

I thank Jasper Rees for help in accessing the GenBank and EMBL nucleic acid data banks.

LITERATURE CITED

BENTLEY, D. L., and T. H. RABBITTS. 1980. Human immunoglobulin variable region genes—DNA sequences of two V_κ genes and a pseudogene. Nature **288**:730–733.

BIRD, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

CLEARY, M. L., E. A. SCHON, and J. B. LINGREL. 1981. Two related pseudogenes are the result of a gene duplication in the goat β-globin locus. Cell **26**:181–190.

COULONDRE, C., J. H. MILLER, P. J. FARABAUGH, and W. GILBERT. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**:775–780.

COX, D. R., and H. D. MILLER. 1965. The theory of stochastic processes. Methuen, London.

DENISON, R. A., S. W. VAN ARSDELL, L. B. BERNSTEIN, and A. M. WEINER. 1981. Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc. Natl. Acad. Sci. USA **78**:810–814.

FLANAGAN, J. G., and T. H. RABBITTS. 1982. Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing gamma, epsilon and alpha genes. Nature **300**:709–713.

GOJOBORI, T., W-H. LI, and D. GRAUR. 1982. Patterns of nucleotide substitutions in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

HARDISON, R. C., E. T. BUTLER, E. LACY, T. MANIATIS, H. ROSENTHAL, and A. EFSTRATIADIS. 1979. The structure and transcription of four linked rabbit β-like globin genes. Cell **18**:1285–1297.

HEINDELL, H. C., A. LIU, G. V. PADDOCK, G. M. STUDNICKA, and W. A. SALSER. 1978. The primary sequence of rabbit α-globin mRNA. Cell **15**:43–54.

HISAJIMA, H., Y. NISHIDA, S. NAKAI, N. TAKAHASHI, S. UEDA, and T. HONJO. 1983. Structure of the human immunoglobulin C-epsilon-2 gene, a truncated pseudogene: implications for its evolutionary origin. Proc. Natl. Acad. Sci. USA **80**:2995–2999.

IKEMURA, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2**:13–34.

JACQ, C., J. R. MILLER, and G. G. BROWNLEE. 1977. A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell **12**:109–120.

KONKEL, D. A., J. V. MAIZEL, and P. LEDER. 1979. The evolution and sequence comparison of two recently diverged mouse chromosomal β-globin genes. Cell **18**:865–873.

LACY, E., and T. MANIATIS. 1980. The nucleotide sequence of a rabbit β-globin pseudogene. Cell **21**:545–553.

LAWN, R. M., A. EFSTRATIADIS, C. O'CONNELL, and T. MANIATIS. 1980. The nucleotide sequence of the human β-globin gene. Cell **21**:647–651.

LEE, M. G.-S., S. A. LEWIS, C. D. WILDE, and N. J. COWAN. 1983. Evolutionary history of a multigene family: an expressed β-tubulin gene and three processed pseudogenes. Cell **33**:477–487.

LI, W.-H., T. GOJOBORI, and M. NEI. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

LI, W.-H., C.-I. WU, and C. C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.

LUND, E., and J. E. DAHLBERG. 1984. True genes for human U1 small nuclear RNA. J. Biol. Chem. **259**:2013–2021.

MCCULLAGH, P., and J. NELDER. 1983. Generalized linear models. Chapman & Hall, London.

MAX, E. E., J. BATTEY, R. NEY, I. R. KIRSCH, and P. LEDER. 1982. Duplication and deletion in the human immunoglobulin ε genes. Cell **29**:691–699.

MICHELSON, A. M., and S. H. ORKIN. 1980. The 3′ untranslated regions of the duplicated human α-globin genes are unexpectedly divergent. Cell **22**:371–377.

MILLER, J. R., E. M. CARTWRIGHT, G. G. BROWNLEE, N. V. FEDOROFF, and D. D. BROWN.

1978. The nucleotide sequence of oocyte 5S DNA in *Xenopus laevis*. II. The GC-rich region. Cell **13**:717–725.

MONSTEIN, H.-J., K. HAMMARSTROEM, G. WESTIN, J. ZABIELSKI, L. PHILIPSON, and U. PETTERSSON. 1983. Loci for human U1 RNA: structural and evolutionary implications. J. Mol. Biol. **167**:245–257.

NISHIOKA, Y., and P. LEDER. 1979. The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell **18**:875–882.

NISHIOKA, Y., and P. LEDER. 1980. Organization and complete sequence of identical embryonic and plasmacytoma κ V-region genes. J. Biol. Chem. **255**:3691–3694.

NISHIOKA, Y., A. LEDER, and P. LEDER. 1980. Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. Proc. Natl. Acad. Sci. USA **77**:2806–2809.

NUSSINOV, R. 1984. Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res. **12**:1749–1763.

PROUDFOOT, N. J., S. GILLAM, M. SMITH, and J. I. LONGLEY. 1977. Nucleotide sequence of the 3′ terminal third of rabbit α-globin messenger RNA: comparison with human α-globin messenger RNA. Cell **11**:807–818.

PROUDFOOT, N. J., and T. MANIATIS. 1980. The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell **21**:537–544.

RAZIN, A., and A. D. RIGGS. 1980. DNA methylation and gene function. Science **210**:604–610.

SCHON, E. A., M. L. CLEARY, J. R. HAYNES, and J. B. LINGREL. 1981. Structure and evolution of goat γ-, β^C-, and α^A-globin genes: three developmentally regulated genes contain inserted elements. Cell **27**:359–369.

SMITH, T. F., W. W. RALPH, M. GOODMAN, and J. CZELUSNIAK. 1985. Codon usage in the vertebrate hemoglobins and its implications. Mol. Biol. Evol. **2**:390–398.

UEDA, S., S. NAKAI, Y. NISHIDA, H. HISAJIMA, and T. HONJO. 1982. Long terminal repeat-like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. EMBO J. **1**:1539–1544.

VAN ARSDELL, S. W., R. A. DENISON, L. B. BERNSTEIN, A. M. WEINER, T. MANSON, and R. F. GESTELAND. 1981. Direct repeats flank three small nuclear RNA pseudogenes in the human genome. Cell **26**:11–17.

MASATOSHI NEI, reviewing editor

Received November 7, 1985; revision received February 3, 1986.